Supplementary Information for

SCREEN: predicting single-cell gene expression perturbation responses via optimal transport

Contents

Supplementary Text S1: Masked variational autoencoder.

The masked variational autoencoder (MVAE) comprises an encoder and a decoder module. The encoder aims at encoding high-dimensional data into a lower-dimensional latent space, assuming Gaussian distribution for the latent embeddings. The decoder is responsible for reconstructing the latent embeddings back to the original data space as faithfully as possible. For a given input gene expression data x, MVAE first simulates the addition of noise by randomly setting a certain proportion of gene expression values to zero. Subsequently, the encoder employs a two-layer feedforward neural network to encode the noise-added data, yielding the corresponding mean vector $\mu(x)$ and variance vector $\Sigma(x)$. In the latent space, the vector z of the corresponding input data is obtained through the method of sampling and reconstruction from the standard normal distribution $N(0,1)$. Specifically, we first sample $\varepsilon \sim N(0,1)$, and then compute $z = \mu(x) + \Sigma^{\frac{1}{2}}(x) \times \varepsilon$. Subsequently, the decoder employs a symmetric two-layer feedforward neural network to further restore the latent vectors to the original data space, resulting in the reconstructed data $\hat{x}$.

The loss function of MVAE consists of two components: one representing the dissimilarity between the original data and the reconstructed data, which is measured using the mean squared error loss; the other capturing the divergence between the encoded latent variable distribution and the standard normal distribution, which is quantified by Kullback-Leibler (KL) divergence. Thus, the loss function can be formulated as:

$$Loss_{DVAE} = MSE(x, \hat{x}) + \alpha KL\left(N\left(\mu(x), \Sigma^{\frac{1}{2}}(x)\right), N(0,1)\right) \quad (1)$$

where α is primarily employed to balance the reconstruction loss and the KL divergence. By updating gradients through back-propagation to minimize the loss function, MVAE simultaneously encourages the generated data to accurately reconstruct input data and urges the latent variable distribution to approach the standard normal distribution.

Supplementary Text S2: Input convex neural networks.

Input convex neural networks (ICNN) are a class of scalar-valued neural networks that enforce the function $x \to f(x;\theta) \in R$ to be convex [1]. Given an input $x \in R^d$, the mapping $x \to f(x;\theta)$ is given by an L-layer feedforward neural network, where the output values at each layer are computed using the following equation:

$$z_{l+1} = \sigma_l(W_l z_l + A_l x + b_l), f(x;\theta) = z_L \quad (2)$$

where $l = 0,1,\cdots,L-1$, $\{W_l\}$ and $\{A_l\}$ are weight matrices (with the convention that $W_0 = 0$), $\{b_l\}$ are bias terms, and σ_l is the activation function at layer l. We represent the set of parameters as $\theta = \{\{W_l\},\{A_l\},\{b_l\}\}$. The function $f(x;\theta)$ is convex when θ satisfies the following conditions:

(i) W_l is a non-negative matrix for $l = 0,1,\cdots,L-1$;

(ii) σ_l is a convex and non-decreasing activation function for $l = 0,1,\cdots,L-1$.

Existing research has shown that any convex function on a compact set can be approximated to the required precision using ICNNs [2]. This suggests that choosing ICNNs as an approximation class for convex functions is feasible.

Supplementary Text S3: Optimal transport mapping.

In this section, we will provide a detailed exposition on deriving the maximin optimization formulation for solving the optimal transport problem and elucidate how to utilize ICNN for its solution.

Let Q and P be two probability distributions on R^d. The Monge optimal transport problem can be defined as follows:

$$\min_{T:T_{\#}Q=P} \quad \frac{1}{2}\mathbb{E}_{X\sim Q} \parallel X - T(X) \parallel^2 \tag{3}$$

For a mapping T: $R^d \to R^d$, $T_{\#}Q = P$ is the measure-preservation condition, that is:

$$(T_{\#}Q)(A) = P(A) = Q\left(T^{-1}(A)\right), \forall A \in \mathcal{B}(R^d) \tag{4}$$

Any mapping T that attains the minimum value in Eq. (3) is called an optimal transport map, which may not exist. The feasible set of the above optimization problem may be empty. To address the issue of the existence of solutions to the Monge problem, we can introduce its relaxed form [3]:

$$W_2^2(P,Q) \triangleq \inf_{\pi\in\Pi(P,Q)} \frac{1}{2}\mathbb{E}_{(X,Y)\sim\pi} \parallel X - Y \parallel^2 \tag{5}$$

where $\Pi(P,Q)$ denotes the set of all joint probability distributions with marginal probability distributions P and Q. Equation (5) is also known as the original formula for the 2-Wasserstein distance. We can also have a corresponding dual formula for Eq. (5), defined as:

$$W_2^2(P,Q) = \sup_{(f,g)\in\Phi_c} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[g(Y)] \tag{6}$$

where Φ_C is the constraint space for the pair of functions (f,g), defined as $\Phi_C = \{(f,g) \in L^1(P) \times L^1(Q): f(x) + g(y) \leq \frac{1}{2}||x-y||_2^2,\ \forall(x,y)\ dP \otimes dQ\ a.e.\}$. Here, $L^1(P) = \{f: f$ is measurable and $\int f\, dP < \infty\}$ represents the set of integrable functions with respect to P.

Villani further simplified Eq. (6) by transforming the optimization problem involving a pair of functions (f,g) to a form that only involves a single convex function f [4]:

$$W_2^2(P,Q) = C_{P,Q} - \inf_{f\in\mathrm{cvx}(P)} \mathbb{E}_P[f(X)] + \mathbb{E}_Q[f^*(Y)] \tag{7}$$

where $CVX(P)$ denotes the set of all convex functions in $L^1(P)$, and $f^*(y) = sup_x\langle x, y\rangle - f(x)$ is the convex conjugate function of f. Villani proved that when the density of Q exists, Eq. (7) has an optimal solution (f, f^*), and ∇f^* is the optimal transport map from Q to P, that is:

$$\mathbb{E}_Q \|\nabla f^*(Y) - Y\|^2 = \inf_{T:T_{\#}Q=P} \mathbb{E}_Q \parallel T(Y) - Y \parallel^2 \tag{8}$$

However, due to the existence of the conjugate function f^*, Eq. (7) is not suitable for standard stochastic optimization schemes. Makkuva et al. proposed a new min-max formula that uses a new convex function to replace the conjugate function [3]:

$$W_2^2(P, Q) = \sup_{f \in \mathrm{cvx}(P)} \inf_{g \in \mathrm{cvx}(Q)} V_{P,Q}(f, g) + C_{P,Q} \tag{9}$$

where

$$V_{P,Q}(f, g) = -\mathbb{E}_P[f(X)] - \mathbb{E}_Q\big[\langle Y, \nabla g(Y)\rangle - f\big(\nabla g(Y)\big)\big] \tag{10}$$

At this point, there exists an optimal pair (f, g), where ∇g represents the optimal transport map from distribution Q to P. For any convex function f, the function $g \in L^1(Q)$ that achieves the lower bound in Eq. (9) is convex and equal to f^*. Therefore, the constraint $g \in CVX(Q)$ can be relaxed to $g \in L^1(Q)$.

In order to obtain a solution to the optimization problem in Eq. (9), Makkuva et al. utilized ICNNs to parameterize functions f and g, and obtained an approximate 2-Wasserstein distance formula as follows [3]:

$$W_2^2(P, Q) = \sup_{f \in \mathrm{ICNN}(\mathbb{R}^d)} \inf_{g \in \mathrm{ICNN}(\mathbb{R}^d)} V_{P,Q}(f, g) + C_{P,Q} \tag{11}$$

We use θ_f and θ_g to represent the parameters of the ICNN. To ensure convexity, all weights W_l in θ_f are strictly constrained to be non-negative, while the constraint is relaxed for θ_g, and a regularization term is introduced as follows:

$$R(\theta_g) = \lambda \sum_{W_l \in \theta_g} \|max(-W_l, 0)\|^2 \tag{12}$$

where $\lambda > 0$ is the coefficient of the regularization term.

In each iteration, samples $\{X_i\}_{i=1}^N$ and $\{Y_i\}_{i=1}^N$ are sampled from distributions P and Q, respectively. The optimization problem corresponding to Eq. (11) is then solved using Eq. (13):

$$max_{\theta_f:W_l\geq 0} min_{\theta_g} \frac{1}{M}\sum_{i=1}^{M}[f\left(\nabla g(Y_i)\right) - \langle Y_i, \nabla g(Y_i)\rangle - f(X_i)] + R(\theta_g) \tag{13}$$

After training, the obtained ∇g is the optimal transport map from distribution Q to distribution P.

Supplementary Text S4: Datasets for model evaluation.

In order to comprehensively assess the performance of SCREEN in different scenarios, we collected seven datasets encompassing various perturbations such as chemical stimulation, bacterial infection, and drug treatment. These datasets cover different cell types and states for prediction purposes. The PBMC dataset consists of human peripheral blood mononuclear cells before and after interferon-β stimulation. The PBMC dataset contains seven cell types and a total of 18,868 cells, including 9,925 unperturbed cells and 8,943 perturbed cells, each with 6,998 genes [5]. Each of the HpolyDay3, HpolyDay10, and Salmonella datasets consists of intestinal epithelial cells of eight cell types [6]. In each dataset, there are 3,240 unperturbed cells but different types of perturbed cells: the HpolyDay3 dataset contains 2,121 cells infected with Heligmosomoides polygyrus for three days, the HpolyDay10 dataset contains 2,711 cells infected with Heligmosomoides polygyrus for ten days, and the Salmonella dataset contains 1,770 cells infected with Salmonella. The Chang dataset consists of 42,275 human non-small cell lung cancer cells, including 21,041 unperturbed cells, 5,920 cells treated with the GNE-69 targeted therapy, 8,737 cells treated with the GNE-104 targeted agent, and 6,577 cells treated with the EGFR inhibitor erlotinib [7]. We divided the Chang dataset into three subsets, namely the Chang_GNE-69, Chang_GNE-104, and Chang_erlotinib datasets, with each set containing only one type of perturbed cells and an equal number of unperturbed cells that were randomly sampled without replacement from all the 21,041 unperturbed cells.

Supplementary Text S5: Detailed description of evaluation metrics.

The first Wasserstein distance (w_1) plays a crucial role in assessing the perturbation-induced changes in the probability distributions of individual genes between the unperturbed state (u) and the perturbed state (v). Specifically, for each gene, denoted by its unique probability distribution, $w_1(u, v)$ is calculated using the formula:

$$w_1(u, v) = \inf_{\pi \in \Gamma(u,v)} \int_{\mathbb{R}\times\mathbb{R}} |x - y| \mathrm{d}\,\pi(x, y)$$

where $\Gamma(u, v)$ signifies the probability distribution set on $\mathbb{R} \times \mathbb{R}$ whose marginals are u and v on the first and second factors respectively. This calculation is systematically employed for each gene. Subsequently, an aggregation is performed, summing the first Wasserstein distances across all the considered genes.

Analogously, for a given gene, the calculation formula of maximum mean discrepancy (MMD) is as follows:

$$MMD[\mathcal{F}, u, v] = \sup_{||f||_{\mathcal{H}} \leq 1} E_u[f(x)] - E_v[f(y)]$$

where $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$. The function space $\mathcal{F}$ is defined as the unit ball in the reproducing kernel Hilbert space $\mathcal{H}$, denoted as $||f|| \leq 1$. Here, $f(x)$ represents the inner product in the reproducing kernel Hilbert space between the vector f and the vector $\phi(x)$, where $\phi(x)$ is the mapping of x to a vector in the Hilbert space through the kernel function. Subsequently, employing a consistent computational approach, we summed MMD values of all the considered genes.

Besides, we calculated the number of common genes between the two sets of differentially expressed genes derived from true and predicted responses, respectively, to evaluate whether the biological variance between the predicted responses and the control group is consistent with the biological variance between the actual responses and the control group. We used Wilcoxon rank-sum tests to identify the set D_p of the top 30 DEGs between the predicted responses and the control group, as well as the set D_t of the top 30 between the true responses and the control group. We then counted the number c of common genes between the two sets of D_p and D_t.

Analogously, we can also consider the top 100 DEGs or the top 50 DEGs instead of the top 30 DEGs.

A smaller MMD or Wasserstein distance indicates a higher similarity between gene expression distribution of the true perturbation responses and gene expression distribution of the predicted responses, and thus means a better prediction performance. A higher c indicates that the biological variance between the predicted responses and the control group is better consistent with the biological variance between the actual responses and the control group, and thus suggests a better performance of predicting single-cell gene expression perturbation responses.

Supplementary Text S6: Model ablation experiments.

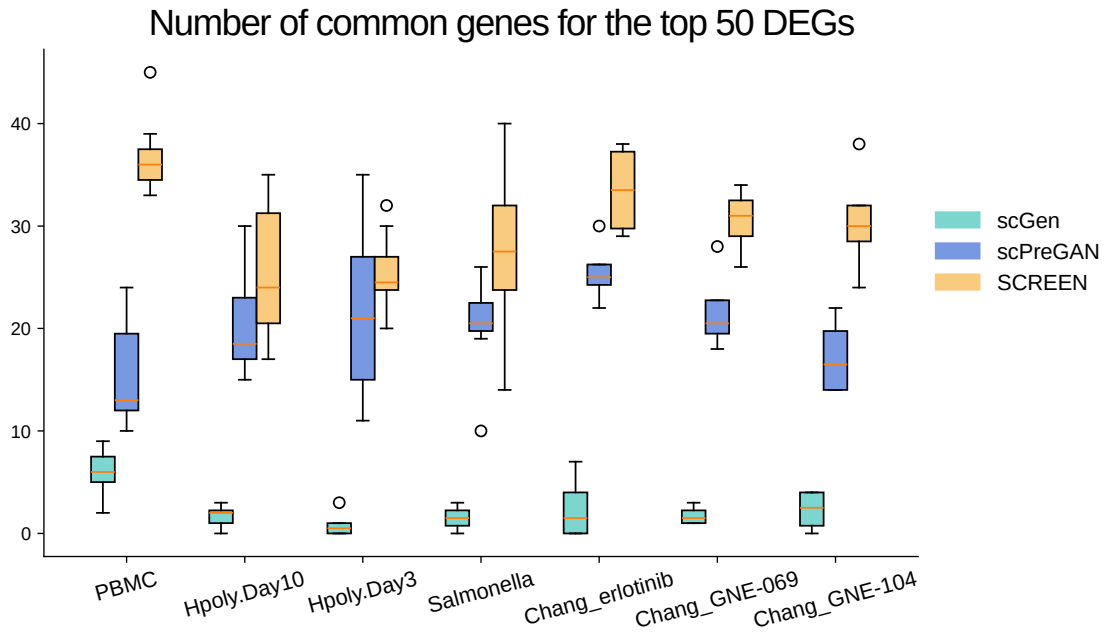
Given that model ablation experiments are important to investigate the contribution of various modules in a model, we replaced our proposed MVAE with the commonly used principal component analysis (PCA). Specifically, we conducted dimensionality reduction using PCA on the original gene expression data, followed by prediction using the optimal transport based on ICNN. This approach is referred to as PCAOT. By comparing the performance of SCREEN and PCAOT, we aim to verify whether MVAE could capture latent features in the original data more effectively. Using the PBMC dataset as an example, we compared the performance of SCREEN and PCAOT via various evaluation metrics. As shown in Supplementary Fig. S2, SCREEN achieved significantly smaller MMD and Wasserstein distance than PCAOT, and significantly outperformed PCAOT in predicting differentially expressed genes, with the former able to identify the majority of the top 30 and the top 100 DEGs, while the latter can identify less than half. Taken together, the results demonstrate that the performance of the SCREEN model surpasses that of the PCAOT model, indicating that MVAE is effective in capturing latent features in single-cell data.

Supplementary Text S7: Imbalance degree of cell types.

The imbalance degree of cell types is employed to gauge whether the sample distribution among distinct cell types within a dataset is uniform. A higher imbalance degree signifies a more uneven distribution of cell types within the dataset. Datasets with higher imbalance degrees typically result in underutilization of information from less abundant cell types, which can affect prediction performance. We utilized entropy to quantify the uncertainty in the distribution of cell types [8]. Given a dataset with C different cell types and the proportion of each cell type denoted as p_i, the cell type imbalance degree I is defined as follows:

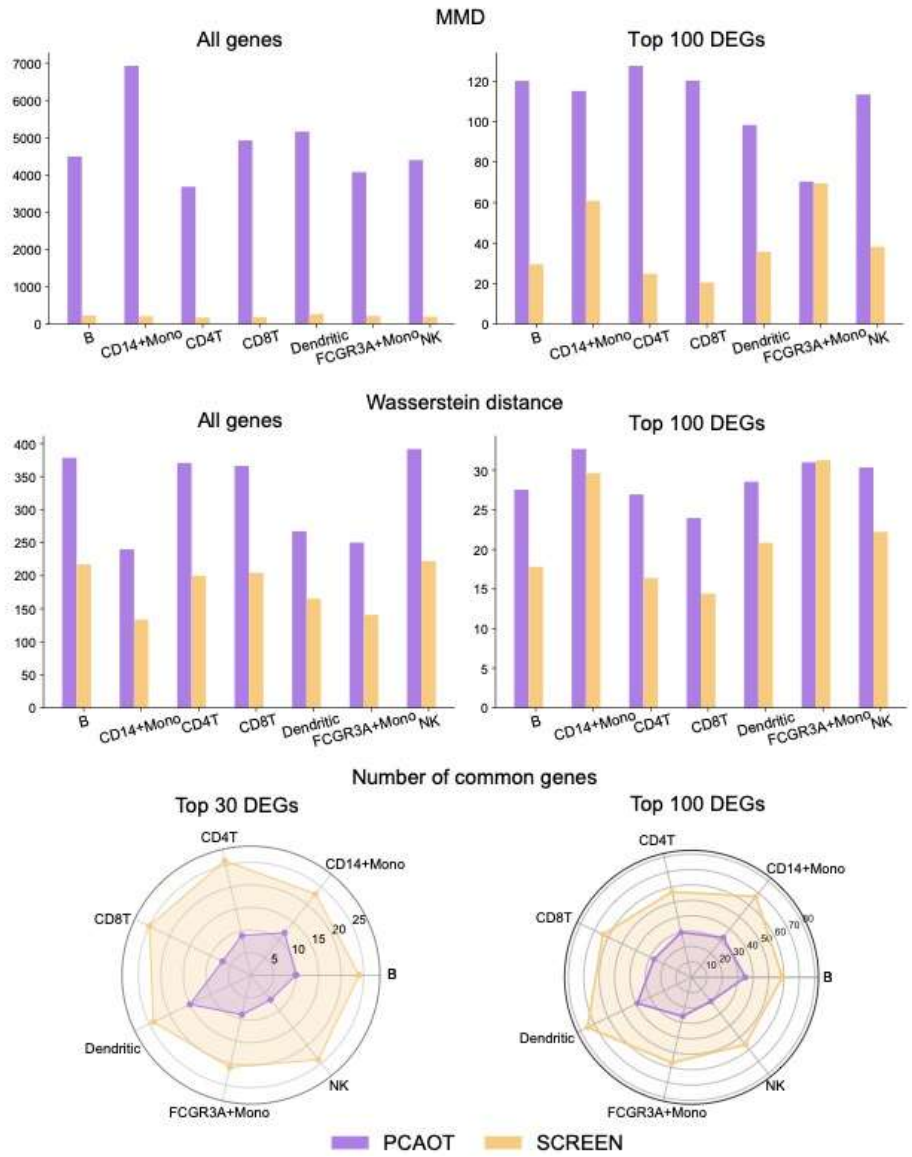
$$I = 1 - \frac{-\sum_{i=1}^{C} p_i \log(p_i)}{\log(C)} \tag{14}$$

Supplementary Fig. S1



Supplementary Fig. S1. Performance of different methods on various datasets evaluated by the number of common genes between the two sets of differentially expressed genes (DEGs) derived from true and predicted responses, respectively.

Supplementary Fig. S2



Supplementary Fig. S2. Performance of SCREEN and PCAOT evaluated by maximum mean discrepancy (MMD), Wasserstein distance, and the number of common genes.

Reference

1. Amos, B., L. Xu, and J.Z. Kolter, *Input Convex Neural Networks*, in *Proceedings of the 34th International Conference on Machine Learning*, P. Doina and T. Yee Whye, Editors. 2017, PMLR: Proceedings of Machine Learning Research. p. 146--155.
2. Chen, Y., Y. Shi, and B. Zhang, *Optimal control via neural networks: A convex approach.* arXiv preprint arXiv:1805.11835, 2018.
3. Makkuva, A.V., et al. *Optimal transport mapping via input convex neural networks*. in *37th International Conference on Machine Learning*. 2020.
4. Villani, C., *Topics in optimal transportation*. Vol. 58. 2021: American Mathematical Soc.
5. Kang, H.M., et al., *Multiplexed droplet single-cell RNA-sequencing using natural genetic variation.* Nature Biotechnology, 2018. **36**(1): p. 89-94.
6. Haber, A.L., et al., *A single-cell survey of the small intestinal epithelium.* Nature, 2017. **551**(7680): p. 333-339.
7. Green, T.D., et al. *scPerturb: Information Resource for Harmonized Single-Cell Perturbation Data*. in *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*. 2022.
8. Chen, S., et al., *ASTER: accurately estimating the number of cell types in single-cell chromatin accessibility data.* Bioinformatics, 2023. **39**(1).